UniFrag and GenomePrimer

van Hijum, SAFT; de Jong, A; Buist, G; Kok, J; Kuipers, OP

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btg203

Publication date:
2003

BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 12 2003, pages 1580–1582
DOI: 10.1093/bioinformatics/btg203

UniFrag and GenomePrimer: selection of primers for genome-wide production of unique amplicons

Sacha A. F. T. van Hijum, Anne de Jong, Girbe Buist, Jan Kok and Oscar P. Kuipers*

Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, PO Box 14, 9750 AA Haren, The Netherlands

Received on January 13, 2003; revised on February 27, 2003; accepted on March 11, 2003

ABSTRACT

Summary: the complementary programs *UniFrag* and *GenomePrimer* were developed to provide a reliable high-throughput method to select the most unique regions within genomic DNA sequence(s) and design primers therein, involving minimal user intervention and maximum flexibility.

Availability: Freely available for educational and research purposes by non-profit institutions at http://molgen.biol.rug.nl/molgen/research/molgensoftware.php

Contact: jonga@biol.rug.nl

Supplementary information: http://molgen.biol.rug.nl/publication/primer_data/

Various programs have recently become available for the selection from coding sequences of single primers (e.g. *OligoArray*, Rouillard *et al.*, 2002, and *ROCK*, Strain and Chmielewski, 2001) or primer pairs (e.g. *PrimeArray*, Raddatz *et al.*, 2001) for the production of DNA-microarrays using oligonucleotides or amplicons. With *GST-PRIME* (Varotto *et al.*, 2001), a large number of primer pairs can be generated starting from a list of accession numbers (GIs). *PrimoUnique* (Chang Biosciences; web-site: http://www.changbioscience.com/primo/primou.html) designs primer pairs for each member of a list of DNA sequences with high similarity (a family). Those primer pairs that might aspecifically amplify a different family member are eliminated.

We use a high-throughput approach for amplicon design and production for DNA-microarrays of bacterial genomes. As some specific features, such as high throughput selection of unique regions and subsequent primer design, were not present in available software packages, the complementary programs *UniFrag* and *GenomePrimer* were developed. Together, these programs allow selecting unique regions within a DNA sequence and automatically designing primers, with minimal user intervention and maximum flexibility. *UniFrag* and *GenomePrimer* were not developed for primer design on genes containing introns. By using unique regions in DNA sequences for primer design, cross-hybridization during DNA microarray experiments is minimized. Also, as aspecific priming during PCR is minimized, PCR is more successful. Moreover, a unified primer design will allow better robotization of PCR.

UniFrag runs on a Unix platform and only requires a locally installed 'Blastall' program (the version used was 2.2.3; http://research.nhgri.nih.gov/blastall/). *UniFrag* consists of five sub-programs written in Pascal and compiled by FreePascal 1.0.4 (http://www.freepascal.org/) under Red Hat Linux release 7.2 (http://www.redhat.com). *GenomePrimer* was written in Borland Delphi 5.0 and runs on any Microsoft Windows platform.

Figure 1 presents a flow scheme of *UniFrag* and *GenomePrimer*. Input files for *UniFrag* are: (i) a FASTA file containing the DNA sequences from which the unique regions have to be selected; (ii) a FASTA file containing a reference set of DNA sequences formatted with *Formatdb* for *Blastall*; and (iii) a configuration file containing the options set by the user. A reference set should consist of, at least, the DNA sequences that are used as input for the *UniFrag* program but other DNA sequences (such as genome sequences) can also be added. In microarray experiments, RNA originating from an organism other than the one represented on the slide might be used. By using the genomic sequence (if available) of the other organism in the reference set of *UniFrag*, possible cross-hybridization will be prevented.

In a typical *UniFrag* run (Fig. 1), overlapping fragments are generated by using a 'window' (a window corresponds to a fragment with a size set by the user) that 'slides' over each of the input sequences. The 'windows' overlap each other with an overlap size that can be set by the user (Fig. 1A and B). By generating overlapping fragments, the chance increases to find a fragment that is unique within a DNA sequence. Sequences that are smaller than the minimum fragment size are saved in a list of 'leftovers'

*To whom correspondence should be addressed.

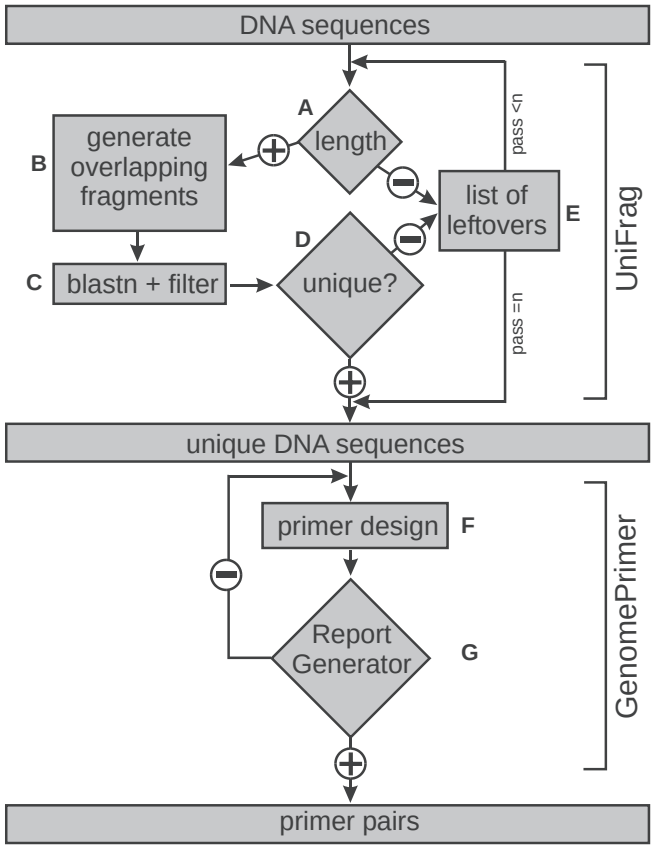

Fig. 1. Flow chart of *UniFrag* and *GenomePrimer*. Arrows with a plus (+) or minus (-) sign indicate data that met or did not meet the selected criteria (in case of steps A and D) or signify a successful or not successful primer design (step G), respectively.

(Fig. 1E). The remaining overlapping DNA fragments are compared using *BlastN* against the reference set and the results are filtered using an expectancy value cutoff parameter (*e*-value) commonly used in blast searches (Fig. 1C). From the filtered blast output, the fragment with the highest expectancy value (most discriminating fragment), is selected for each input DNA sequence (Fig. 1D). If the selected fragment has an *e*-value higher than the 'cutoff unique' parameter, the input DNA sequence is used for primer design; if not, the input DNA sequence is saved in the 'leftover' list (Fig. 1E). *UniFrag* can be run for any number of passes (*n* in Fig. 1), decreasing fragment and overlap sizes until the maximum number of unique fragments is obtained (Table 1, supplementary information). Other sequences can be added to the resulting list of fragments (i.e. sequences of the 'leftover' list). The FASTA file generated can then be used for subsequent *GenomePrimer* processing.

Via user-friendly interface frames, various selection criteria can be chosen before running the *GenomePrimer* design program, which selects thousands of primers within seconds (Fig. 1F). Primers are examined for: (i) equal distribution of G and C; (ii) occurrence of palindromic sequences; and (iii) homology between primers of a primer pair. The melting temperature can be set according to either one of two rules: (i) $T_m = 4 \times (GC) + 2 \times (AT)$ (Suggs *et al.*, 1981); or (ii) $T_m = 62.3 + 0.41 \times (GC) - (500/\text{length})$ (Sugimoto *et al.*, 1996). Desired tags, for instance for the re-amplification of all amplicons using a tag-specific oligonucleotide pair, can be easily introduced. General result statistics such as success rate, average primer length, number of nucleotides to be synthesized and specific characteristics such as the number of (too short and/or too low GC-content) amplicons are displayed by the Report Generator (Fig. 1G). If primer design failed in too many cases, the settings can be easily adjusted and a new design can be performed until all primers/amplicons meet the desired characteristics. Those DNA sequences on which no primer pair could be designed by the *GenomePrimer* selection criteria can be selected for primer design using other criteria.

The freely available web-based version of *PrimoUnique* allows designing one unique primer pair at a time. In contrast to *UniFrag* and *GenomePrimer*, *PrimoUnique* does not allow constraints in amplicon length. As only the primers and not the amplicon sequences are unique, *PrimoUnique* is not usable for effective genome-wide primer design in contrast to the combination of *UniFrag* and *GenomePrimer*. The *PrimeArray* program allows using a predetermined amplicon length cut-off whereas in *GenomePrimer* a window of amplicon sizes is used. The variation in amplicon size allows primers to be chosen more flexibly in a location on the DNA sequence. This allows using more stringent primer characteristics during the design, which results in a more efficient (and successful) high-throughput approach because subsequent PCR conditions can be standardized. *GenomePrimer* provides the user with a number of criteria for primer design: (i) preferred location of primers in a sequence (for instance to obtain 3′-, 5′- or central amplicons of a gene); (ii) preferred primer length and reduction of primer length, while maintaining annealing properties, to reduce production costs; (iii) 3′- G or C on primers to improve the PCR extension step, which is especially important for low GC organisms such as *Lactococcus lactis*; and (iv) preferred GC content of the primer.

UniFrag was used to select the unique regions in all predicted open reading frames (ORFs) of *L.lactis* IL1403 and *Streptococcus pneumoniae* TIGR4 (Table 1, supplementary information). More unique ORF fragments could be selected from *L.lactis* IL1403 (1516; 68% of the total amount of 2214 ORFs) than from *S.pneumoniae* TIGR4 (1390; 56% of the total amount of 2495 ORFs) because the genome sequence of the latter contained more

ORFs smaller than 500 bp, which were ignored in the analysis. The *S.pneumoniae* TIGR4 genome contained less ORFs (69; 2.8%) in which no unique fragment could be identified than *L.lactis* IL1403 (80; 3.6%). In most cases, the ORFs of which no unique fragment could be identified were specified by insertion elements or transposon sequences. Table 1 (supplementary information) clearly illustrates that decreasing the fragment size results in a significantly higher amount of unique ORF fragments. By additionally decreasing the overlap size, the probability increases that a certain fragment is unique.

GenomePrimer was initially tested on the predicted ORFs of the *L.lactis* bacteriophage r1t (Table 2, supplementary information). A 100% success rate for primer design and amplicon production of the 51 r1t ORFs was obtained. For the selection of primers for ORFs larger than 90 bp in the complete genomes of *S.pneumoniae* TIGR4 and *L.lactis* IL1403, a smaller range of amplicon length was used (80–800 bp) than for r1t (89–1842 bp). Success rates of both primer sets (100% and 99.5% for *L.lactis* and *S.pneumoniae* TIGR4, respectively) were high in the first round of PCR. PCR reactions were carried out under standard conditions in a Bio-Rad iCycler 1 PCR machine (Bio-Rad, Hercules, CA) in 96-wells format with a non-proofreading DNA polymerase.

Reliable high-throughput genome-wide primer design on unique fragments of ORFs is the strength of using *UniFrag* and *GenomePrimer* software. Using unique fragments in microarray studies should reduce cross-hybridizations, which will improve data quality and make data analysis more straightforward.

REFERENCES

Raddatz,G., Dehio,M., Meyer,T.F. and Dehio,C. (2001) PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, **17**, 98–99.

Rouillard,J.M., Herbert,C.J. and Zuker,M. (2002) OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, **18**, 486–487.

Strain,S.R. and Chmielewski,J.G. (2001) ROCK: a spreadsheet-based program for the generation and analysis of random oligonucleotide primers used in PCR. *Biotechniques*, **30**, 1286–1291.

Suggs,S.V., Wallace,R.B., Hirose,T., Kawashima,E.H. and Itakura,K. (1981) Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta 2-microglobulin. *Proc. Natl Acad. Sci. USA*, **78**, 6613–6617.

Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, **24**, 4501–4505.

Varotto,C., Richly,E., Salamini,F. and Leister,D. (2001) GST-PRIME: a genome-wide primer design software for the generation of gene sequence tags. *Nucleic Acids Res.*, **29**, 4373–4377.